OPIC
OFFICE DE LA PROPRIÉTÉ INTELLECTUELLE DU CANADA

CIPO
CANADIAN INTELLECTUAL PROPERTY OFFICE

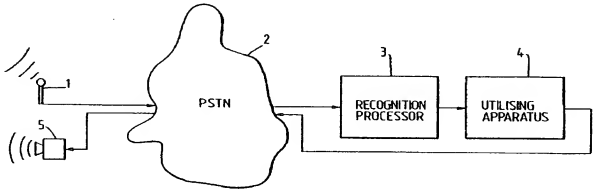
(12) (19) (CA) **Brevet-Patent**

(11) (21) (C) **2,158,847**
(86) 1994/03/25
(87) 1994/09/29
(45) 2000/03/14

(72) Pawlewski, Mark, GB
(72) Tang, Joseph Gordon, GB
(73) British Telecommunications public limited company, GB

(51) Int.Cl.[6] G10L 5/06
(30) 1993/03/25 (93302302.0) EP

(54) **METHODE ET APPAREIL DE RECONNAISSANCE VOCALE**
(54) **A METHOD AND APPARATUS FOR SPEAKER RECOGNITION**



(57) Dispositif servant à identifier un locuteur et comprenant des moyens (210, 220, 230) servant à générer, en réaction à un signal vocal, une pluralité de données de caractéristiques comprenant une série d'ensembles de coefficients, chaque ensemble comprenant une pluralité de coefficients indiquant l'amplitude spectrale sur une courte durée dans une pluralité de bandes de fréquence, ainsi que des moyens (260) servant à comparer lesdites données de caractéristiques à des données prédéterminées de référence de locuteurs, ainsi qu'à indiquer l'identification d'un locuteur en fonction de ladite comparaison; le dispositif est caractérisé par le fait que lesdites bandes de fréquence sont espacées irrégulièrement le long de l'axe de fréquence et par des moyens (250) servant à calculer une grandeur spectrale moyenne sur une longue durée d'au moins un desdits coefficients, ainsi qu'à normaliser ledit coefficient ou chacun desdits coefficients par l'intermédiaire de ladite moyenne calculée sur une longue durée.

(57) Apparatus for speaker recognition which comprises means (210, 220, 230) for generating, in response to a speech signal, a plurality of feature data comprising a series of coefficient sets, each set comprising a plurality of coefficients indicating the short term spectral amplitude in a plurality of frequency bands, and means (260) for comparing said feature data with predetermined speaker reference data, and for indicating recognition of a corresponding speaker in dependence upon said comparison; characterised in that said frequency bands are unevenly spaced along the frequency axis, and by means (250) for deriving a long term average spectral magnitude of at least one of said coefficients; and for normalising the or each of said at least one coefficient by said long term average.

2158847

PCT

WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 5 :
G10L 5/06

A1

(11) International Publication Number: WO 94/22132

(43) International Publication Date: 29 September 1994 (29.09.94)

(21) International Application Number: PCT/GB94/00629

(22) International Filing Date: 25 March 1994 (25.03.94)

(30) Priority Data:
93302302.0 25 March 1993 (25.03.93) EP
(34) Countries for which the regional or international application was filed: AT et al.

(60) Parent Application or Grant
(63) Related by Continuation
US 08/105,583 (CIP)
Filed on 13 August 1993 (13.08.93)

(71) Applicant *(for all designated States except US)*: BRITISH TELECOMMUNICATIONS PUBLIC LIMITED COMPANY [GB/GB]; 81 Newgate Street, London EC1A 7AJ (GB).

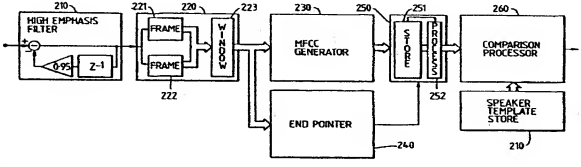
(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: PAWLEWSKI, Mark [GB/GB]; 35 Beverley Road, Ipswich, Suffolk IP4 4BU (GB). TANG, Joseph, Gordon [GB/GB]; 200 Richmon Road, Kingston Upon Thames, Surrey KT2 5HE (GB).

(74) Agent: ROBERTS, Simon, Christopher; BT Group Legal Services, Intellectual Property Dept., 13th floor, 151 Gower Street, London WC1E 6BA (GB).

(81) Designated States: AU, CA, CN, FI, JP, KR, NZ, US, European patent (AT, BE, CH, DE, DK, ES, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).

Published
With international search report.
Before the expiration of the time limit for amending the claims and to be republished in the event of the receipt of amendments.

(54) Title: A METHOD AND APPARATUS FOR SPEAKER RECOGNITION



(57) Abstract

Apparatus for speaker recognition which comprises means (210, 220, 230) for generating, in response to a speech signal, a plurality of feature data comprising a series of coefficient sets, each set comprising a plurality of coefficients indicating the short term spectral amplitude in a plurality of frequency bands, and means (260) for comparing said feature data with predetermined speaker reference data, and for indicating recognition of a corresponding speaker in dependence upon said comparison; characterised in that said frequency bands are unevenly spaced along the frequency axis, and by means (250) for deriving a long term average spectral magnitude of at least one of said coefficients; and for normalising the or each of said at least one coefficient by said long term average.

A METHOD AND APPARATUS FOR SPEAKER RECOGNITION

The present invention relates to speech processing and in particular to processing for speaker recognition. Recognition processing includes speaker recognition, in which the identity of the speaker is detected or verified, and speech recognition, in which a particular word (or, sometimes, a phrase or a phoneme, or other spoken matter) is detected. Speech recognition includes so-called speaker-independent recognition, in which speech data derived from multiple speakers is used in recognition processing, and so-called speaker dependant recognition, in which speech data derived from a single speaker is used in recognition processing. In general, in speech recognition, the processing aims to reduce the effects on the spoken word of different speakers, whereas in speaker recognition the reverse is true.

It is common in recognition processing to input speech data, typically in digital form, to a so-called front-end processor, which derives from the stream of input speech data a more compact, more perceptually significant set of data referred to as a front-end feature set or vector. For example, speech is typically input via a microphone, sampled, digitised, segmented into frames of length 10-20ms (e.g. sampled at 8 KHz) and, for each frame, a set of K coefficients (typically 5-25) is calculated. Since there are N frames e.g. 25-100 per word, there are N x K (in the order of 1,000) coefficients in a feature vector. In speaker recognition the speaker to be recognised is generally assumed to be speaking a predetermined word, known to the recognition apparatus and to the speaker (e.g. a PIN in banking). A stored representation of the word, known as a template, comprises a reference feature matrix of that word previously derived from a speaker known to be genuine. The input feature matrix derived from the speaker to be recognised is compared with the template and a measure of similarity

between the two is compared with a threshold for an acceptance decision.

A problem arises from the tendency of speakers to vary the rate at which words are spoken, so that an input speech matrix corresponding to a given word may be longer (i.e. consist of more frames) or shorter than the template for that word. It is therefore necessary for the recognition apparatus to time-align the two matrices before a comparison can be made, and one well known method of time-alignment and comparison is the Dynamic Time Warp (DTW) method described, for example, in "Speaker Independent Recognition of words using Clustering Techniques", Rabiner etal, IEEE Trans. on ASSP, vol 24, no. 4, August 1979.

Various features have been used or proposed for recognition processing. In general, since the features used for speech recognition are intended to distinguish one word from another without sensitivity to the speaker whereas those for speaker recognition are intended to distinguish speakers for a known word or words, a feature suitable for one type of recognition may be unsuitable for the other. Some features for speaker recognition are described in "Automatic Recognition of Speakers from their voices", Atal, Proc IEEE vol 64 pp 460-475, April 1976.

One known type of feature coefficient is the cepstrum. Cepstra are formed by performing a spectral decomposition (e.g. a spectral transform such as the Fourier Transform), taking the logarithm of the transform coefficients, and performing an inverse spectral decomposition.

In speaker recognition, the LPC (Linear Prediction Coefficient) cepstrum and FFT (Fast Fourier Transform) cepstrum features are known, the former being more widely used.

In speech recognition, a known feature is the mel-frequency cepstrum coefficient (MFCC). A description of an algorithm for calculating MFCC's, and calculating a distance measure between an MFCC feature vector and a word template using Dynamic Time Warping is given in "On the evaluation of

Speech Recognisers and Data Bases using a Reference System", Chollet & Gagnoulet, 1982 IEEE, International Conference on Acoustics, Speech and Signal Processing, pp2026-2029, incorporated herein in its entirely (including its references).

An MFCC feature vector in general is derived by performing a spectral transform (e.g. a FFT), on each frame of a speech signal, to derive a signal spectrum; integrating the terms of the spectrum into a series of broad bands, which are distributed in an uneven, so-called 'mel-frequency' scale along the frequency axis; taking the logarithms of the magnitude in each band; and then performing a further transform (e.g. a Discrete Cosine Transform (DCT)) to generate the MFCC coefficient set for the frame. It is found that the useful information is generally confined to the lower order coefficients. The mel-frequency scale may, for example, be frequency bands evenly spaced on a linear frequency scale between 0-1 Khz, and evenly spaced on a logarithmic frequency scale above 1 KHz.

MFCC's eliminate pitch information, which is useful for speech recognition since this varies between speakers, but undesirable for speaker recognition. MFCC's have accordingly not been preferred for speaker recognition.

In "Speaker Recognition by Statistical Features and Dynamic Features", Review of Electrical Communications Laboratories, Vol 30, No 3, 1982, PP 467-482, S Furui describes and compares four speaker recognition methods. In two of the methods, the spectral envelope is represented by log area rations, and in the other two by LPC-cepstrum coefficients. In the case of the use of LPC-cepstrum coefficients, the coefficients are averaged over the duration of the entire utterance, which may be several words (eg up to 7), and the averaged values are subtracted from the cepstrum coefficients of every frame to compensate for frequency - response distortions introduced by the transmission system. Time functions for the normalised cepstrum coefficients are then expanded by an orthogonal polynomial representation over

90ms intervals every 10 ms. The polynomial processing gives rise to a polynomial representation which corresponds to the mean value, slope and curvature of each cepstrum coefficient in each segment. The original time functions of the cepstrum coefficients are used in place of the zeroth order polynomial coefficients, together with the first and second-order polynomial coefficients, so that each utterance is represented by a time function of a 30-dimensional vector. From these 30 elements, a set of 18 elements is chosen with a view to expanding the overall distance distribution of customer and imposter sample utterances (determined during an extended training/enrolment phase). In an alternative processing arrangement, LPC analysis is followed by a statistical feature extraction process which involves the generation of, inter alia, Fourier cosine expansion coefficients. With this alternative processing, the final feature set consists of 60 selected elements. This latter processing arrangement ("Method 2") which used statistical features including Fourier coefficients extracted from time functions of LPC cepstrum coefficients and fundamental frequency, was reported to produce a 99.9 recognition accuracy on the particular training and imposter data used. Furui concluded that the results "indicate that LPC cepstrum coefficients are much more efficient than log area ratios". Furui provides no teaching to use any other cepstra, nor to filter the input speech into a plurality of predetermined frequency bands. Moreover, Furui implicitly teaches the value of using a very large feature set (eg up to 60 elements) - something which it is clearly desirable to avoid, particularly if the relevant recognizer population is very large.

In "Speaker Verification over Long Distance Telephone Lines", ICASSP 89, Vol 1, 23 May 1989, Pages 524-527, J M Naik et al compare speaker verification techniques using either template-based Dynamic Time Warping or Hidden Markov Modelling. Again, LPC analysis was performed to provide the pre-processed information from which features were extracted.

32 parameters were initially computed from each frame of LPC data:

* speech level estimate in dB;
* RMS frame energy in dB;
* a scalar measure of rate of spectral change;
* 14 filter-bank magnitudes in dB
  - mel-spaced simulated filter banks
  - normalised by frame energy
* time difference of frame energy over 40ms;
* time difference of 14 filter bank magnitudes over 40ms.

The speech features used to compare the reference and test templates were obtained by a linear transformation of these 32 parameters, to produce a set of 18 features for use as a template. No explanation or reason is given for the use of the mel-spaced filter banks and there is absolutely no suggestion to use any other form of cepstra than LPC.

It is an object of the invention to provide a method and apparatus for speaker recognition, using an improved front end feature.

Accordingly, the invention provides a method and apparatus for speaker recognition in which a speech signal is processed to derive recognition features comprising a plurality of spectral amplitude coefficients, distributed on an uneven frequency scale, at least one of which is normalised by its average level over a time interval.

The normalisation acts to reduce the effect of any long term filtering of the speech signal by, for example, a telecommunications channel through which it passes.

For example, the coefficients may be distributed on the mel-frequency scale. In this case they may be Mel-frequency Cepstral Coefficients. The normalisation is conveniently performed by forming the long term arithmetic mean value of each coefficient, and subtracting this from each coefficient value; since the coefficients are logarithmic, the subtraction corresponds to division by the geometric mean of the source signal in each mel-frequency

band.

It is found that mel-frequency cepstral coefficients, when normalised in this manner, provide a reliable feature set for recognising speakers, using only a relatively small number of coefficients, and moreover, that their use is particularly suited to telecommunications applications because the dependence upon the characteristics of the transmission channel between the speaker and the recognition apparatus is significantly reduced.

Other aspects and preferred embodiments of the invention are as disclosed and claimed herein, with advantages that will be apparent hereafter.

The invention will now be described by way of example only, with reference to the following description and drawings, in which:

Figure 1 shows schematically the employment of a recognition processor according to the invention in a telecommunications environment;

Figure 2 is a block diagram showing schematically the functional elements of a recognition processor according to an embodiment of the invention;

Figure 3 is a flow diagram showing schematically the operation of an MFCC generator forming part of Figure 2;

Figure 4 is an illustrative diagram in the frequency domain illustrating part of the process of Figure 3;

Figure 5 is a flow diagram showing in greater detail the operation of an end point detector forming part of Figure 2;

Figure 6 is an illustrative diagram of amplitude against time to illustrate the process of Figure 5;

Figure 7 is a flow diagram indicating schematically the operation of a normalisation processor forming part of Figure 2 in a preferred embodiment;

Figure 8a is a flow diagram illustrating the operation of a comparison processor forming part of Figure 2 in a speaker verification embodiment;

Figure 8b is a flow diagram illustrating the operation

of a comparison processor forming part of Figure 2 in a speaker identification embodiment;

Figure 9 is a flow diagram showing the operation of a normalisation processor forming part of Figure 2 in an alternative embodiment to that of Figure 7;

Figure 10a is an illustrative plot of MFCC coefficient values against time for each of two different telecommunications channels; and

Figure 10b is a corresponding plot of coefficients normalised according to the preferred embodiment of Figure 7.

PREFERRED EMBODIMENT

Referring to Figure 1, a telecommunications system including speaker recognition generally comprises a microphone, 1, typically forming part of a telephone handset, a telecommunications network (typically a public switched telecommunications network (PSTN)) 2, a recognition processor 3, connected to receive a voice signal from the network 2, and a utilising apparatus 4 connected to the recognition processor 3 and arranged to receive therefrom a voice recognition signal, indicating recognition or otherwise of a particular speaker, and to take action in response thereto. For example, the utilising apparatus 4 may be a remotely operated banking terminal for effecting banking transactions.

In many cases, the utilising apparatus 4 will generate an auditory response to the speaker, transmitted via the network 2 to a loudspeaker 5 typically forming a part of the subscriber handset.

In operation, a speaker speaks into the microphone 1 and an analog speech signal is transmitted from the microphone 1 into the network 2 to the recognition processor 3, where the speech signal is analysed and a signal indicating identification or otherwise of a particular speaker is generated and transmitted to the utilising apparatus 4, which then takes appropriate action in the event of recognition of the speaker.

Typically, the recognition processor needs to acquire data concerning the identity of the speaker against which to verify the speech signal, and this data acquisition may be performed by the recognition processor in a second mode of operation in which the recognition processor 3 is not connected to the utilising apparatus 4, but receives a speech signal from the microphone 1 to form the recognition data for that speaker. However, other methods of acquiring the speaker recognition data are also possible; for example, speaker recognition data may be held on a card carried by the speaker and insertable into a card reader, from which the data is read and transmitted through the PSTN to the recognition processor prior to transmission of the speech signal.

Typically, the recognition processor 3 is ignorant of the route taken by the signal from the microphone 1 to and through the network 2; the microphone 1 may, for instance be connected through a mobile analog or digital radio link to a network 2, or may originate from another country, or from one of a wide variety of types and qualities of receiver handset. Likewise, within the network 2, any one of a large variety of transmission paths may be taken, including radio links, analog and digital paths and so on. Accordingly, the speech signal Y reaching the recognition processor 3 corresponds to the speech signal S received at the microphone 1, convolved with the transfer characteristics of the microphone 1, link to network 2, channel through the network 2, and link to the recognition processor 3, which may be lumped and designated by a single transfer characteristic H.

Recognition Processor 3

In Figure 2, the functional elements of a recogniser according to the preferred embodiment are shown. A high emphasis filter 210 receives the digitised speech waveform at, for example, a sampling rate of 8 KHz as a sequence of 8-bit numbers and performs a high emphasis filtering process (for example by executing a $1 - 0.95.z^{-1}$ filter), to increase

the amplitude of higher frequencies. A speech frame generator 220 receives the filtered signal and forms a sequence of frames of successive samples. For example, the frames may each comprise 256 contiguous samples, and each frame may be overlapped with the succeeding and preceding frames by 50%, so as to give frames of length 32ms, at a rate of 1 per 16ms. For example, a pair of frame buffers 221, 222, with a 16ms differential delay, may be filled in parallel and read out in alternation.

To eliminate spurious frequency artifacts due to the discontinuities at the start and end of each frame, preferably each frame is then passed through a Hamming window processor 223, which (as is well known) scales down the samples towards the edge of each window.

Each frame of 256 windowed samples is then processed by an MFCC generator 230 to extract a set of MFCC coefficients (for example 8 coefficients). At the same time, each windowed frame is supplied to an end point detector 240 which detects the start and finish of a speech utterance and supplies a speech/non-speech control signal to a normaliser 250 comprising a coefficient store memory 251 and a normalising processor 252. The normaliser 250, after receiving a 'speech start' signal from the end pointer 240, stores some or all of the 8 coefficients for each successive frame in the normaliser coefficient store 251 until the 'end of speech' signal is received from the end pointer 240. At this point, the normalising processor 252 calculates from the stored coefficients for each speech frame in store 251 an arithmetic mean coefficient value for each of the 8 coefficients. The arithmetic mean coefficient value for each coefficient is then subtracted from the respective stored coefficient value for each frame, to provide a normalised matrix comprising 8 x N coefficients (where N is the number of frames between the start point and the end point of a spoken utterance).

This normalised coefficient matrix is supplied to a comparison processor 260, which reads a corresponding matrix

associated with a given speaker from a speaker template 270, performs a comparison between the two, and generates a recognition/non-recognition output signal in dependence upon the similarity between the normalised speech vector and the speaker template from the speaker template store 270.

The high emphasis filter 210, window processor 223, MFCC generator 230, end pointer 240, normalising processor 252 and comparison processor 260 may be provided by one or more digital signal processor (DSP) devices and/or microprocessors, suitably programmed, with the frame buffers 221, 222, coefficient store 251 and speaker template store 270 provided within read/write memory devices connected thereto.

MFCC Generation

Referring to Figure 3, the process performed by the MFCC generator 230 comprises performing a Fourier transform on each frame, to provide 256 transform coefficients, in a step 401; forming the power spectrum of the speech signal from the Fourier coefficients by summing the squares of the Real and Imaginary components at each frequency, to provide a 128 coefficient power spectrum in step 402; integrating the power spectrum over 19 frequency bands in a step 403 as discussed in greater detail below with reference to Figure 4, to provide 19 band power coefficients; taking the log (for example to base 10) of each of the 19 coefficients in a step 404; performing a discrete cosine transform on the 19 log values in a step 405, and selecting the lowest order 8 coefficients in a step 406.

The discrete cosine transform is well known and described in, for example, the above referenced Chollet and Gagnoulet paper; briefly, the Nth cosine component of M_m is given by

$$Y_m = \frac{\sum_{n=1}^{N} A(n) \cdot Cos(m\Pi(n+0.5))}{N}$$

where N is the

number of discrete frequency bands (in this case 20, with a frequency domain rotation applied to obtain the 20th point) and A(n) is the amplitude in the mth frequency band. The effect of the DCT is to decorrelate the coefficients A(n) to a large extent.

Referring to Figure 4, Figure 4a notionally indicates a portion of the power spectrum produced in step 402. Figure 4b shows a corresponding portion of the mel-frequency triangular integration windows along the frequency axis. The triangular windows comprise ten windows linearly spaced along the frequency axis, each overlapping its neighbours by 50%, between 0 - 1KHz and a further ten windows, triangular and overlapping by 50% on a logarithmic frequency scale above 1 KHz.

Figure 4c shows schematically the effect of multiplying, pointwise, each sample in the power spectrum by the corresponding term in one of the triangular windows; for clarity, only even number windows have been shown.

Next, the windowed values of Figure 4c are integrated across each window, to provide a single summed coefficient corresponding to that window, as shown in Figure 4d.

The 19 coefficients thus produced (the zero frequency coefficient M_0 being ignored) thus correspond to the power which would be generated in the output of each of a corresponding set of band pass filters, filters below 1 KHz having equal evenly spread band widths and those above 1 KHz having band widths which are equal and evenly spread on a logarithmic frequency scale.

Endpointing

Referring to Figures 5 and 6, the operation of the end pointer 240 of Figure 2 will now be discussed in greater detail.

The endpointer 240 initially squares and sums the signal values within each frame to provide a measure P of the power or energy in the frame.

The value of P is tested against the first threshold P_L, which is set at a relatively low level such that it may

occasionally be crossed by noise signals. No action is taken until a frame has a power level above this low threshold P_L. On a frame rising above the low threshold P_L, a flag indicating a number assigned to that frame is stored (shown as a variable "START" in Fig. 5).

When the value of the power P in a frame rises above an upper threshold P_H, which corresponds to the presence of speech and which is at a level above likely noise levels, speech is assumed to be present. The point taken as the start point of the speech is a frame of predetermined number ("LEAD") of frames before that ("START") at which the signal rose above the low threshold P_L. In this way, although speech is only confirmed to be present when the signal level rises above the high threshold, the start of the utterance is not lost. Accordingly, the number of the frame thus calculated as the start point is output by the endpointer 240 to control the normaliser 250.

If the level of the speech signal remains between the two thresholds for longer than a predetermined time T_{max}, the value "START" is cleared.

On the frame energy dropping from the upper threshold P_H below the lower threshold P_L, the end pointer 240 waits through a predetermined number of frames termed the "overhang" time T_{oh}. If the level rises above the lower threshold P_L again, within the overhang time, speech is assumed still to be present. Once the power level of the signal has fallen below the lower threshold P_L for more than T_{oh} frames, the utterance is assumed to be over, and the endpointer outputs an end point frame number which corresponds to current frame number, less the number of frames T_{oh} (i.e. the point at which the signal was last at the threshold P_L), plus a predetermined number of frames termed the "LAG".

Normalisation

Referring to Figure 7, the normalisation process carried out by the normaliser 250 will now be described in greater detail.

The frames of 8 MFCC coefficients per frame are stored in the coefficient buffer 251 successively. After the endpointer 240 detects the end of the spoken utterance, it signals the start and end frame numbers to the normaliser 250. The normalising processor 252 then, for each of the 8 coefficients, recalls the value of that coefficient from the memory for all frames between the start and the end frame and forms the arithmetic mean by adding the coefficient values and dividing by N, the number of frames between the start and the end frames. This provides a set of 8 average values $\bar{M}_i$ ($i = 1$ to 8).

Next, for each coefficient of each frame, the normalising processor 252 calculates a normalised coefficient value $G_{i,k}$, (where K is the frame number) by subtracting the corresponding average value $\bar{M}_i$ from each coefficient value $M_{i,k}$.

The set of 8 x N coefficients making up the normalised vector $G_{i,k}$ are then output by the normalisation processor 252.

Comparison Processing

A detailed description of the comparison processor 260 is unnecessary since its operation is conventional. Figure 8a indicates schematically the operation of the comparison processor in speaker verification; in essence the comparison processor reads the feature vector G comprising the normalised MFCCs; reads a speaker template T comprising a corresponding reference vector of coefficients; performs a comparison between the two vectors using for example the well known Dynamic Time Warp algorithm to time-align the two (using, for example, the algorithm given in the above Chollet and Gagnoulet paper) to generate a scalar distance measure D indicating the difference between the two vectors and tests the distance measure D against the threshold. If the distance D is lower than the threshold, the speaker is accepted as corresponding to the stored template; otherwise the speaker is rejected. Figure 8b shows the corresponding operation of the comparison processor 260 in speaker

identification; in this case, a plurality of different vectors T_i are read in succession from the template store 270, and the speech vector G is compared with each in turn to generate a corresponding distance metric D_i. The speaker is then identified as corresponding to the template from which the speech vector differs the least (i.e. which gave rise to the smallest metric D_i).

ALTERNATIVE EMBODIMENTS

In the foregoing embodiment, as discussed in relation to Figure 7, the normaliser 250 needs to know both the start point and the end point of the utterance before it can calculate N (the number of frames between the start point and the end point), and the sum of the coefficient values M_{TOT}, and hence the average value of each coefficient, and hence the normalisedvalue of each coefficient. Accordingly, the normaliser 250 must await detection of the end point by the endpointer 240, and subsequent recognition processing is delayed until the end of the utterance. In many applications, and with fast hardware, this delay may not give rise to difficulties. However, in other applications it may be preferable to begin normalisation before the end of the utterance.

Accordingly, in a first alternative embodiment, instead of normalising the coefficients by subtracting the arithmetic mean value of each coefficient over the whole utterance, coefficients are normalised by subtracting a running average which is updated periodically (for example, on a frame by frame basis).

Referring to Figure 9, accordingly, in this embodiment after the endpointer 240 signals the beginning of the utterance, the normalisation processor 252 reads, for each coefficient, the present average value for that coefficient $\bar{M}_i$; subtracts this from the value M_i of the MFCC coefficient to form a normalised coefficient G_i; increments a frame counter N; adds the coefficient value $\bar{M}_i$ to the current total value M_{TOT}, and divides the sum by the frame counter N, the result being stored as the new value of the coefficient

average value $\bar{M}_i$. The normalised coefficient values G_i for each frame can therefore be released immediately.

It is anticipated that a running average of this type is likely to perform slightly less well than the preferred embodiment, since initially the "average" value is not formed from a representative number of samples. However, some improvement in the performance is nonetheless anticipated when compared to unnormalised coefficients. Naturally, other methods of calculating a running average (for example, using a moving window of past samples or updating less frequently than every frame) are equally possible. In embodiments of this type, the coefficient store 251 may be dispensed with.

In the foregoing embodiments, description has been made of endpointing and normalising over a single contiguous utterance (i.e. a single word). If speaker identification based on several separate words is to be performed, then the process described in the above embodiments could be repeated for each successive word in isolation. However, some information useful in discriminating speakers may be found in the relative level of the coefficient values of each word relative to the others.

Accordingly, in a further embodiment the long term average value $\bar{M}_i$ formed by the normaliser 250 is formed over all the words of the utterance. In a first example according to this embodiment, this is achieved by forming the average over all the frames between the start and end points of each word, as if the words followed immediately one after another as a single utterance, and ignoring the non-speech frames in between the words.

The same result is achieved in a second example, by deriving separate averages as in the foregoing embodiments, for each word, and then adding the averages each weighted by the respective number of frames in the word from which it is derived, so as to form a weighted average from all the words, and then dividing each coefficient of every word by the weighted average derived across all words.

In the foregoing two examples, the weight given to the

average corresponding to each individual word varies depending upon the length of the word, which in turn varies with the rate at which the speaker speaks the word (which is variable independently of the spectral characteristics of the way in which the speaker speaks the word).

Accordingly, in an alternative embodiment, a long term average is formed by forming the average over each word in isolation as before, and then forming a weighted average from the individual averages, but employing predetermined weights corresponding, for example, to the length of the corresponding stored template in the template store which represents that word, rather than the actual duration of the word as in the previous examples. In this way, the dependence on the rate at which the words are spoken is reduced.

Although it might under some circumstances be possible to dispense with the end pointer 240 and form a long term average over the entire duration of the telephone call, in practice this is generally not preferred because during periods of non-speech, the received signal level is generally too low to provide a reliable indication of the spectrum of the communication channel and, furthermore, it is difficult to separate the spectrum of the channel from that of any noise present.

In the above described embodiments, recognition processing apparatus suitable to be coupled to a telecommunications exchange has been described. However, in another embodiment, the invention may be embodied on simple apparatus connected to a conventional subscriber station connected to the telephone network; in this case, analog to digital conversion means are provided for digitising the incoming analog telephone signal.

Although reference is made to the use of programmable digital signal processing (DSP) devices, it will equally be recognised that a conventional general purpose microprocessor operating at sufficient speed might instead be employed. Likewise, a custom designed large scale integration (LSI)

logic circuit could be employed.

The invention has been described with reference to MFCC's, but filter banks on uneven frequency scales approximating to, or differing from, the mel-frequency scale could be used. Whilst triangular windows have been described above for the integration in the frequency domain, it will be appreciated that other window shapes could equally be employed. Whilst a digital processor for calculating MFCC values has been described, it would in principle be possible to provide instead a plurality of band pass analog or digital filters, corresponding to the bands shown in Figure 5b, and to sample the power in each filter band.

Whilst the invention has been shown to be surprisingly advantageous in application to MFCC's, its application to other front end features (preferably Cepstral features) is not excluded.

Whilst a comparison process using the Dynamic Time Warp (DTW) process has been discussed, the invention is equally applicable to recognition employing other types of comparison processing. For example, comparison processing employing hidden Markov modelling (HMM), as disclosed in British Telecom Technology Journal, Vol. 6, No. 2 April 1988, S.J. Cox "Hidden Markov Models for Automatic Speech Recognition : Theory And Application", pages 105-115, or Neural Networks (for example of the well known multilayer perceptron (MLP), or the "self-organising" types, both of which are discussed in the same issue of the British Telecom Technology Journal) may be used.

Whilst the application of the invention to speaker recognition has herein been described, it will be apparent that aspects of the invention may also be applicable to other recognition tasks (e.g. speech recognition).

TEMPLATE GENERATION

In general, the present invention employs a stored reference model ("template" for DTW recognition) for the or each speaker to be identified. Methods of deriving reference

models are well known, and for understanding the present invention it is therefore sufficient to indicate that each template may be formed by a process of inputting a plurality of utterances of the same word by a speaker; digitising the utterances; deriving the normalised set of coefficients G_i in the same way as discussed above for each of the utterances; aligning the utterances in time using, for example, a Dynamic Time Warp process; and then averaging the time aligned coefficient vectors of the utterances to derive an averaged coefficient vector which provides the reference model T. In other words, the process of forming a reference model for use with a given feature set in subsequent recognition is generally the same as the process of deriving the feature set itself, a number of feature sets being averaged to give the reference model.

EFFECTS OF THE INVENTION

Referring to Figure 10, Figure 10a (the left-hand column) shows for each of the 8 MFCCs a graph of coefficient value over time during an utterance. In each case, two traces are shown; these correspond to the same recorded utterance transmitted via two different transmission channels. It will be noted that, particularly in the second and seventh coefficient, the channel results in a substantially constant offset between the two traces, corresponding to the difference in transfer characteristic in the corresponding frequency bands between the two channels.

In the Dynamic Time Warp process, as in other processes in which portions of two patterns to be compared are brought into time alignment, the Dynamic Time Warp process essentially shifts portions of a waveform along the time axis to find a match with another waveform. Where, as here, two waveforms are vertically displaced, then this process of shifting along the time axis (i.e. horizontal shifting) will result in a mismatch and hence in increased likelihood of misrecognition or reduced likelihood of correct recognition.

Referring to Figure 10b, in the right-hand column, the corresponding plots of normalised MFCC's according to the invention are shown. By referring to the 2nd, 6th and 7th coefficients in particular, it will be seen that removal of the average value has in each case brought the two traces into closer alignment. Thus, when a speech vector is compared with a template which may have been derived through a different communications channel, the Dynamic Time Warp comparison processing is less likely to misidentify the speaker due to the effect of the transmission channel.

As noted above, the (generally linear) path from the speaker to the recognition processor can be represented by a lumped transfer characteristic H, comprising the product of cascaded transfer functions of successive stages of the path. Thus, in the frequency domain, each spectral component of the speech signal received by the recognition processor comprises the product of the spectral component of the voice of the speaker with the corresponding spectral component of the transfer function of the communication channel or path. Thus, if the transfer characteristic H of the channel were known the effect of the channel on the speech signal could be removed by dividing each term of the received signal spectrum by the corresponding term of the transfer characteristic H.

However, in a telecommunications system, because of the plurality of diverse alternative signal paths it is not possible directly to model the channel transfer function H. However, it is observed that the channel transfer function is generally spectrally stationary (i.e. does not change much over time). Accordingly, if a time series of a single spectral component is examined, the transfer function acts as a constant multiplicative factor on each value in the series. The geometric mean of each component in the time series is therefore the product of this constant factor and the geometric mean of the original time series. Thus, the effect of the channel is eliminated if each term in the received speech signal spectrum is divided by its long term average.

In taking logarithms of each spectral term, rather than forming the long term geometric mean and dividing thereby it is possible to form the long term arithmetic mean of the logged spectral term and subtract this arithmetic mean from each spectral term.

In the windowing and integrating stages of generating MFCCs, it is thought that there is some transformation of this relationship, so that the foregoing analysis does not entirely apply to the normalisation of MFCCs but is merely illustrative of the effect of the invention.

The normalisation process removes not only the effect of the channel, but also some speech and speaker information. It might therefore be thought that this would reduce the accuracy of recognition, by removing data which could be used to distinguish between two speakers. In fact, surprisingly, after extensive experiments we have found that this is not the case.

CLAIMS

1. A method of speaker recognition, comprising deriving recognition feature data from an input speech signal, said recognition feature data comprising a plurality of coefficients each related to the speech signal magnitude in a predetermined frequency band; comparing said feature data with predetermined speaker reference data; and indicating recognition of a speaker in dependence upon the comparison; characterised in that said frequency bands are unevenly spaced along the frequency axis and in that the step of generating said coefficients includes a step of deriving a long term average spectral magnitude; and processing at least one of said coefficients so as to generate a normalised coefficient in which the effect of said long term magnitude is substantially reduced.

2. A method according to claim 1, in which the frequency bands are spaced on a mel-frequency scale.

3. A method according to claim 1, in which the frequency bands are spaced linearly with frequency below a predetermined limit and logarithmically with frequency above said limit.

4. A method according to any of claims 1 to 3 in which the step of generating said coefficients includes a step of generating a logarithm of said magnitude, generating a logarithmic long term average value and subtracting the logarithmic long term average from the logarithmic magnitude.

5. A method according to any of claims 1 to 4, in which said comparison is such as to time-align the feature data with the reference data.

6. A method according to claim 5, in which the comparison

employs a Dynamic Time Warp process.

7. A method according to any of claims 1 to 6, further comprising the step of recognising a speech start point and a speech end point within said input speech signal; and of deriving said long term average over the duration between said start point and said end point.

8. A method according to any of claims 1 to 7, in which said long term average comprises the long term mean.

9. A method according to any of claims 1 to 7, in which said long term average comprises a moving average which is periodically updated.

10. A method according to any preceding claim, comprising inputting a plurality of words one after another, and forming said long term average over all of said words.

11. Apparatus for speaker recognition which comprises means (210, 220, 230) for generating, in response to a speech signal, a plurality of feature data comprising a series of coefficient sets, each set comprising a plurality of coefficients indicating the short term spectral magnitude in a plurality of frequency bands, and means (260) for comparing said feature data with predetermined speaker reference data, and for indicating recognition of a corresponding speaker in dependence upon said comparison; characterised in that said frequency bands are unevenly spaced along the frequency axis, and by means (250) for deriving a long term average spectral magnitude of at least one of said coefficients; and for normalising the or each of said at least one coefficient by said long term average.

12. Apparatus according to claim 11, in which the frequency bands are spaced on a mel-frequency scale.

13. Apparatus according to claim 11, in which the frequency bands are spaced linearly with frequency below a predetermined limit and logarithmically with frequency above said limit.

14. Apparatus according to any of claims 11 to 13 in which the means (230) for generating said coefficients are arranged to generate a logarithm of said magnitude, generate a logarithmic long term average value and subtract the logarithmic long term average from the logarithmic coefficient magnitude.

15. Apparatus according to any of claims 11 to 14, in which said comparison means (260) is arranged to time-align the feature data with the reference data.

16. Apparatus according claim 15, in which the comparison means (260) employs a Dynamic Time Warp process.

17. Apparatus according to any of claims 11 to 16, further comprising means (240) for recognising a start point and an end point within said speech signal, in which said normalising means (250) is arranged to derive said long term average over the duration of the utterance between said start point and said end point.

18. Apparatus according to any of claims 11 to 17, in which said long term average comprises the long term mean.

19. Apparatus according to any of claims 11 to 17, in which said long term average comprises a moving average which is periodically updated.

20. Apparatus according to any of claims 11 to 19, arranged for inputting a plurality of words one after another, in which said normalising means (250) is arranged to form said long term average over all of said words.

21. Apparatus according to any of claims 11 to 20 adapted to be connected to a telephone network.

22. A telephone network comprising apparatus according to claim 21.

23. Apparatus for recognition processing of a voice signal, comprising means (210, 220, 230) for deriving recognition data comprising a plurality of signals each related to the short term amplitude in a corresponding frequency band of said voice signal, and means (260) for performing recognition processing in dependence thereon; characterised by means (250) for periodically generating or updating a moving long term average spectral amplitude in said frequency bands, and for processing said feature data using said long term average to reduce their dependence upon stationary spectral envelope components.

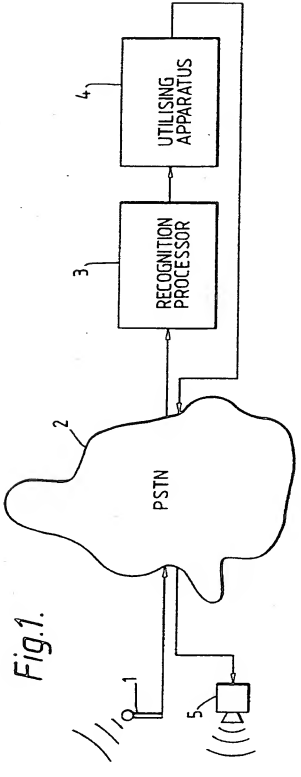

PSTN
RECOGNITION PROCESSOR
UTILISING APPARATUS
1
2
3
4
5

Fig.1.

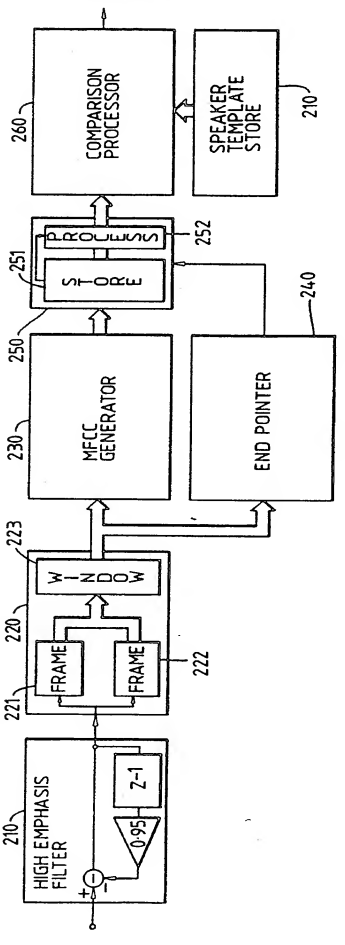

HIGH EMPHASIS FILTER
Z-1
0.95
FRAME
FRAME
WINDOW
MFCC GENERATOR
END POINTER
STORE
PROCESS
COMPARISON PROCESSOR
SPEAKER TEMPLATE STORE
210
220
221
222
223
230
240
250
251
252
260

Fig.2.

Fig.3.

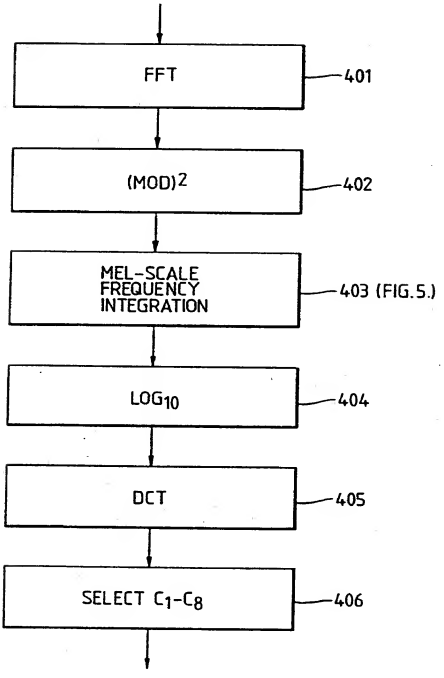

FFT
401
(MOD)2
402
MEL-SCALE FREQUENCY INTEGRATION
403 (FIG.5.)
LOG10
404
DCT
405
SELECT C1-C8
406

Fig.4.

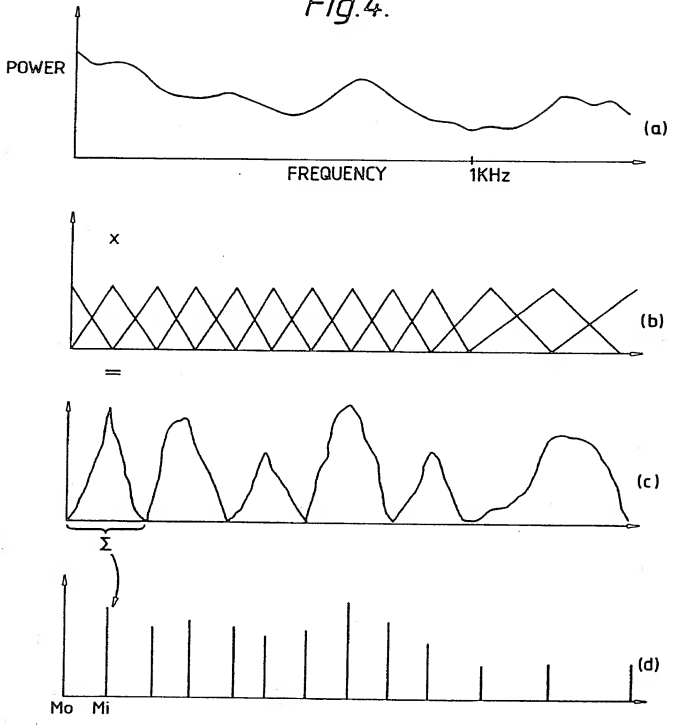


Fig.5.

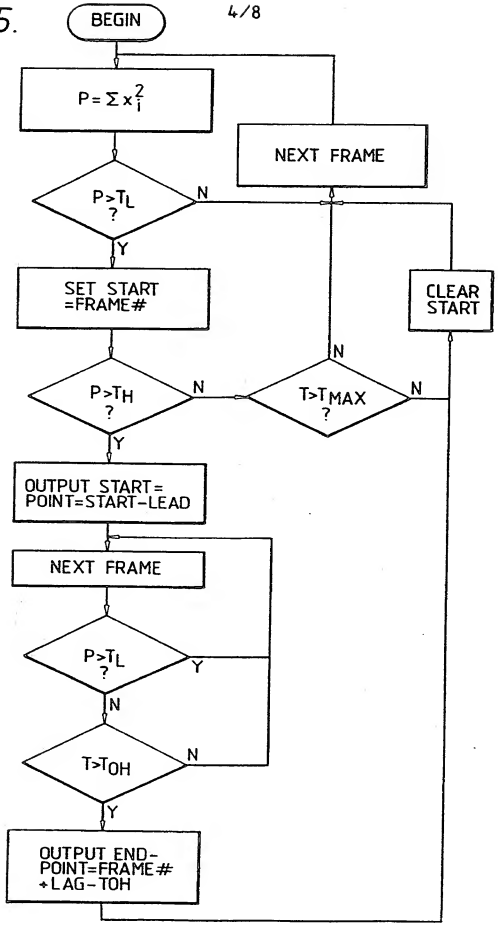

BEGIN
P=Σx_i^2
NEXT FRAME
P>T_L ?
N
Y
SET START =FRAME#
CLEAR START
N
P>T_H ?
N
T>T_MAX ?
N
Y
OUTPUT START= POINT=START-LEAD
NEXT FRAME
P>T_L ?
Y
N
T>T_OH
N
Y
OUTPUT END-POINT=FRAME# +LAG-TOH

Fig. 6.

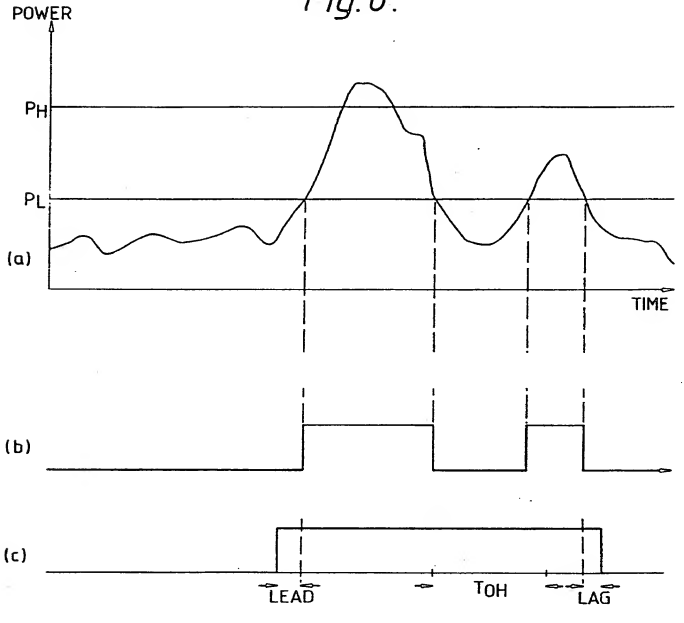


Fig. 7.

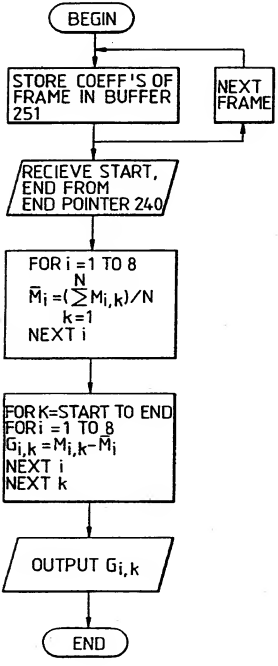

BEGIN
STORE COEFF'S OF FRAME IN BUFFER 251
NEXT FRAME
RECIEVE START, END FROM END POINTER 240
FOR i =1 TO 8
$\bar{M}_i = (\sum_{k=1}^{N} M_{i,k})/N$
NEXT i
FOR K=START TO END
FOR i =1 TO 8
$G_{i,k} = M_{i,k} - \bar{M}_i$
NEXT i
NEXT k
OUTPUT $G_{i,k}$
END

Fig. 9.

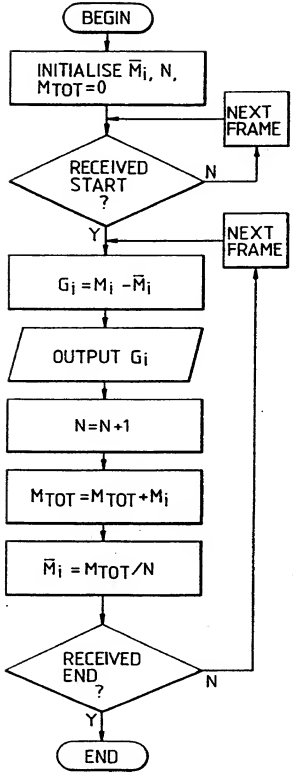

BEGIN
INITIALISE $\bar{M}_i$, N, $M_{TOT}=0$
NEXT FRAME
RECEIVED START ?
N
Y
NEXT FRAME
$G_i = M_i - \bar{M}_i$
OUTPUT G_i
N=N+1
$M_{TOT} = M_{TOT} + M_i$
$\bar{M}_i = M_{TOT}/N$
RECEIVED END ?
N
Y
END

Fig. 8a.

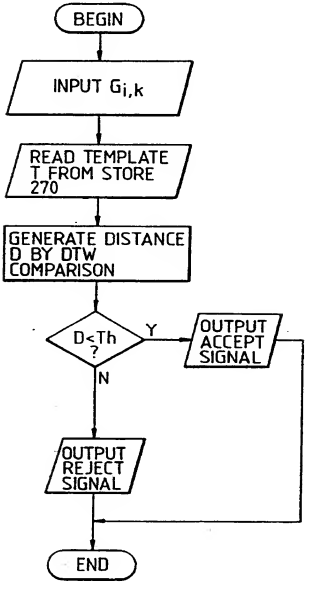

BEGIN
INPUT $G_{i,k}$
READ TEMPLATE T FROM STORE 270
GENERATE DISTANCE D BY DTW COMPARISON
D<Th ?
Y
OUTPUT ACCEPT SIGNAL
N
OUTPUT REJECT SIGNAL
END

Fig. 8b.

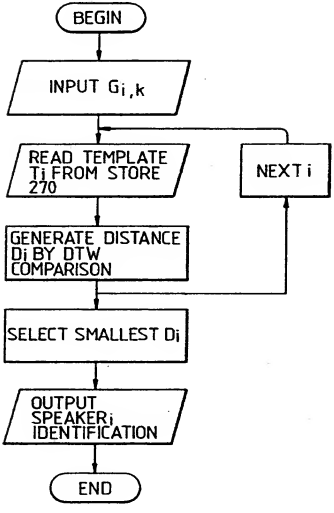
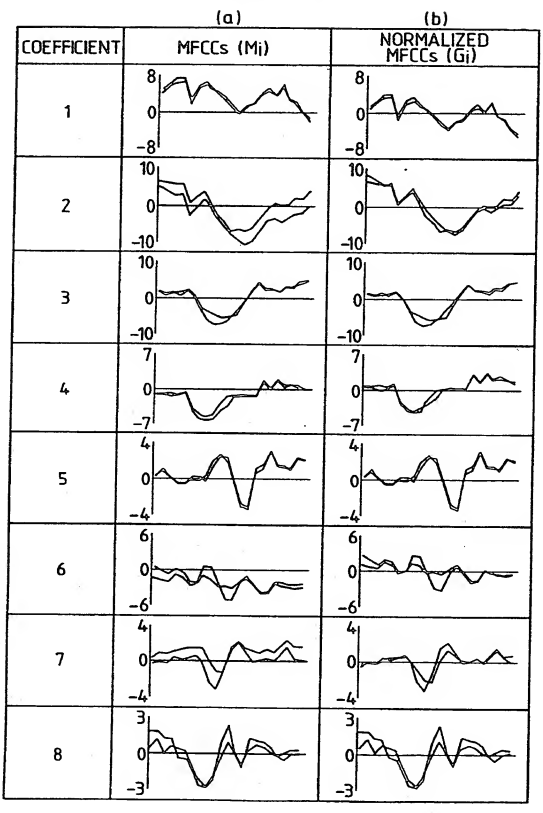

BEGIN
INPUT $G_{i,k}$
READ TEMPLATE Ti FROM STORE 270
NEXT i
GENERATE DISTANCE Di BY DTW COMPARISON
SELECT SMALLEST Di
OUTPUT SPEAKERi IDENTIFICATION
END

Fig.10.


(a)
(b)
COEFFICIENT
MFCCs (Mi)
NORMALIZED MFCCs (Gi)
1
2
3
4
5
6
7
8
8
0
-8
10
0
-10
7
0
-7
4
0
-4
6
0
-6
3
0
-3